Contributed Session 1A
Tuesday, January 23, 2001
11:00 – 12:30

A Bayesian Analysis of the NHANES III BMI Data With Nonignorability

Balgobin Nandram (1), Jai Won Choi (2), and Myron Katzoff (2)
Department of Mathematical Sciences, Worcester Polytechnic Institute (1)
Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention (2)

OBJECTIVES: We investigate the efficacy of employing the approach of a continuous model expansion on an odds ratio to model nonignorable nonresponse. This is considered for the situation of studying the health status of the US population with data collected from the National Health and Nutrition Examination Survey (NHANES).

METHODS: We consider data that can be modeled using the normal means model and we assume the data are collected for a survey in which there is significant nonresponse. We propose a method to incorporate uncertainty about ignorability of the nonreponse mechanism by centering a nonignorable model on an ignorable one. We apply our method to the NHANES data to study body mass index (BMI), one indicator of the health status of the US population. We incorporate covariates such as age on a continuous scale, and race and sex, which are used to explain the variation in BMI data. The implementation requires computational effort of a type that is accomplished with Monte Carlo Markov Chain methods.

RESULTS: We predict the finite population mean BMI in each of 34 counties (population 500,000) in the NHANES survey. In addition, we describe how to assess the goodness of fit of the model using cross validation. We also show how BMI is related to the covariates mentioned above.

CONCLUSION: We have studied nonignorable nonresponse for inferences on BMI means through a full Bayesian approach. We have addressed uncertainty about ignorability through a parameter that centers a nonignorable model on an ignorable one. This way of examining the data (i.e., using the centering parameter) allows the investigator to characterize the variability that arises from uncertainty about ignorability.

Correlation Coefficient Estimation When Both Variables Are Left Censored: Application to HIV Viral Load Assay Data

Robert H. Lyles and Jovonne K. Williams
Department of Biostatistics, Rollins School of Public Health, Emory University

OBJECTIVES: For many current HIV studies, assessing the level of correlation between two different measures of viral load obtained from each of a sample of patients is one of the goals. This assessment has important consequences, including the potential to provide information regarding likelihood of disease transmission, and potential implications for cost effectiveness when a less invasive or expensive assay is sought. A complication for the analyst seeking valid point and confidence interval estimates of the correlation in such a setting is the fact that both variables are likely subject to some level of left censoring due to values below assay detection limits. Our purpose is to investigate a parametric approach based upon a bivariate normal likelihood accounting for left censoring of two variables that may have different detection limits.

METHODS: By determining the contributions of each of the four possible types of paired observations that can arise, we derive the appropriate likelihood. We provide simulation results to evaluate sampling properties of the resulting correlation estimator in finite samples. In an effort to provide improved confidence interval coverage over that obtained using the standard Wald-type approach, we also evaluate profile likelihood-based intervals. We apply the approach to actual HIV viral load data from a CDC-sponsored study, and we describe an extension of the original model to incorporate interval censoring as motivated by the study design.

RESULTS: Empirical results over a variety of conditions suggest that the maximum likelihood estimator of the correlation displays minimal bias, and that Wald-type confidence intervals provide slightly subnominal coverage. More limited results indicate that the profile likelihood intervals give improved coverage at the expense of being slightly wider. Ad hoc estimators based upon eliminating or inserting fixed values for non-detects are shown to be seriously biased.

CONCLUSION: In situations where Pearson's coefficient would be the appropriate measure of correlation but is unavailable due to left censoring of both viral load assay variables in question, the likelihood presented can provide a means for obtaining valid point and confidence interval estimates of the true correlation.

Inference with Conditional Mean Imputation and with Applications to the Analysis of the National Survey of Lead and Allergens in Housing

Ming Yin (1), Richard Cohn (1), Patrick J. Vojta (2), Michael L. Muilenberg (3), Harriet Burge (3), Warren Friedman (4), and Darryl Zeldin (2)
Analytical Sciences, Inc. (1)
National Institute of Environmental Health Sciences (2)
Harvard University School of Public Health (3)
Department of Housing and Urban Development (4)

OBJECTIVES: The National Survey of Lead and Allergen in Housing is a multi-stage complex sample survey sponsored by the Department of Housing and Urban Development (HUD) and National Institute of Environmental Health Sciences (NIEHS) to assess children's potential household exposure to lead and allergens.

One of the analysis goals is to derive the national percentage estimates of households with elevated allergen level above some critical points, e.g., concentration level of cockroach allergen (Bla g I) >= 0.1 U/gm. However, two kinds of complications are involved in this analysis. First, some of the lab measurements are completely missing due to insufficient or no dust samples collected for the lab assay. Second, the measurement is subject to lower detection limits in the laboratory, and these limits vary from one sample to another. Complication is introduced when an important threshold cut off point is coincident with one of the smaller values of lower detection limits. Failure to account for these two kinds of missing data in the analysis may result in significant biases in the percentage estimates and their corresponding confidence intervals. The objective of this presentation is to develop methods to properly account for this characteristic of the data in estimating percentages and their confidence intervals.

METHODS: Imputation methods remain attractive for handling missing data. Little and Rubin (1987) discussed various imputation strategies. More recently, Schafer and Schenker (2000, JASA) proposed an analytic method to produce appropriate variance estimates for conditional predictive mean imputation. One limitation for their method is that it assumes that missings are ignorable, so their method cannot be applied directly to our setting where both ignorable and non-ignorable missing data are present.

In this paper, we expand their method to a more general setting including both types of missing data, where the non-ignorable missing mechanism is known up to a few parameters. This method is based on asymptotic expansions of point estimators and their associated variance estimators and produces a first order approximation to Rubin's repeated-imputation inference with an infinite number of imputations. It can be more efficient than multiple imputation with a small number of imputations. Next, we apply this newly developed analytic method to our allergen survey analysis.

RESULTS: National percentage estimates of households with elevated allergen levels and their confidence intervals are derived using our newly developed method, a multiple imputation method and the naïve method (ignoring all missing data). Survey weights and design information are incorporated into the analysis. Both the new method and the multiple imputation method show substantial corrections for percentage estimates and variance estimates from those derived using the naïve method. Also, we conduct simulation studies to contrast the new method and multiple imputation method with variable sizes of imputations (M) and with different kinds of missing mechanisms.

CONCLUSION: We expand Schafer and Schenker's analytic method to a more general setting, and apply it successfully to the analysis of the National Survey of Lead and Allergen in Housing. Through empirical studies, we show that the newly developed method can be more efficient than multiple imputation with a small number of imputations.

An Alternative Method for Forming Nonresponse Adjustment Cells, Applied to Children Without Vaccination Histories from Providers in the National Immunization Survey

Philip J. Smith (1), David C. Hoaglin (2), Michael P. Battaglia (2), J.N.K. Rao (3), and Danni Daniels (4)
Data Management Division, National Immunization Program, Centers for Disease Control and Prevention (1),
Abt Associates Inc. (2)
School of Mathematics & Statistics, Carleton University (3)
Division of HIV/AIDS Prevention, National Center for HIV, STD, and TB Prevention, Centers for Disease Control and Prevention (4)

OBJECTIVES: In 1998, provider verified vaccination histories were obtained from approximately 70% of all children sampled in the National Immunization Survey (NIS). Statistical analyses that fail to adjust NIS sampling weights for differences between children who have provider data and children who do not have provider data may yield misleading inferences. This paper presents statistical methods that evaluate the extent of provider vaccination history nonresponse and makes appropriate adjustments to survey weights when this bias is detected.

METHODS: In this paper we present an alternative method for forming cells to adjust for unit nonresponse; it is often effective to group sample units according to either their response propensities or their predictive probabilities. To combine these two approaches, we use response propensity as the basis for an initial set of cells and then use predictive probability to refine each cell (separately). A straightforward algorithm determines the total number of adjustment cells and the extent of refinement, including the possibility of basing the cells entirely on response propensity or entirely on predictive probability.

RESULTS: For the 1998 samples in the 78 Immunization Action Plan areas of the NIS, two or three adjustment cells, based on predictive probability, are generally best, in the sense of bias reduction.

CONCLUSION: The results suggest that the NIS encounters little nonresponse bias from the provider-reported vaccination histories that it cannot obtain.

Contributed Session 1B
Tuesday, January 23, 2001
11:00 – 12:30

Hierarchical Modeling For Serial Dilution Assays

Andrew Gelman (1), Ginger Chew (2), and Wenbin Lu (1)
Department of Statistics, Columbia University (1)
Division of Environmental Health Sciences, Columbia University (2)

OBJECTIVES: The standard analysis of serial dilution assays results in a high fraction of missing data, with many or even most of the observations recorded as above or below detection limit. Our objective is to take advantage of the partial information in these measurements--not merely treating them as missing or even censored, but rather using the noisy but somewhat informative numerical measurements. The goal is to obtain more precise assays, especially for measurements of low concentrations.

METHODS: We fit a hierarchical model to standard reference curve and assay data, allowing a common parametric form but with variance between dilutions and between plates. The model includes both dilution and assay errors and is thus a form of nonlinear regression with errors in both x and y. We use Bayesian inference to estimate the unknown concentrations for assay data; the model automatically allows for larger errors at the extreme range, and there is no need to discard data as above or below detection limits.

RESULTS: We fit the model to data from assays of cockroach allergens Bla g1 and Bla g2 in dust samples, in an experiment in which about half of the measurements were deemed above or below detection limit using standard methods. We also apply similar models to assess the effects of unintended thawing and refreezing of frozen liquified dust samples. In addition to the estimates of allergen concentrations, we suggest the possibilities of new designs that allow more effective pooling of standard reference data.

CONCLUSION: We model assay data hierarchically with errors in assays, dilutions, and plate variability, building on the existing literature of models for serial diultion data (e.g., Higgins, Davidian, Chew, and Burge, 1998, and Racine-Poon, Weihs, and Smith, 1991). The between- plate variance component allows us to partially pool information, including standard reference curves, and make use of the information available in very low and high measurements.

A Hierarchical Model for Antimicrobial Assay Analysis

Alex R. Varbanov and Charles H. Taylor
Health Care Research Center, Procter & Gamble Co.

OBJECTIVES: Quantitative antimicrobial assays are used to assess the efficacy of chemical germicides. Standard methods for their analysis rely on log transformation of the response, the count of viable microbes, to justify use of normal theory statistical inference. The log reduction (LR), the difference on the log scale between average surviving microbes for control and test carriers, is used as an efficacy parameter. That parameter is not on the original response scale, which complicates its interpretation. The presence of two different definitions of LR makes the statistical inference even more difficult. In addition, the current statistical methods for antimicrobial assay analysis rely on normal distribution asymptotic theory, which might not work well for small samples. To overcome those problems, a new Bayesian approach is introduced.

METHODS: A hierarchical model with three levels is used to describe antimicrobial assay data. The first level explains variability between the observations from the same carrier. It introduces a survival fraction of microbial count for each treatment to evaluate its efficacy. The second level of the model incorporates information about a typical control count of microbes. Random effects are used to model additional variability within and between carriers. Prior distribution on some nuisance parameters is specified at the third model level. A Gibbs Sampling Algorithm is used to generate samples from the posterior distribution of the parameters. Statistical inference about typical treatment counts or survival fractions can be easily made based on those samples.

RESULTS: An example is used to illustrate the powerful statistical inference allowed by the new Bayesian approach. The posterior distribution of each parameter in the model can be estimated from the generated samples. For example, some posterior quantiles for the survival fraction are given. Inference about functions of parameters is done by evaluating the function at the generated parameter values. Probability intervals for typical treatment microbial counts and differences of them are given to evaluate efficacy and compare treatments on the response scale.

CONCLUSION: The proposed Bayesian approach does not rely on normal distribution asymptotic theory to get standard errors or interval estimates for the parameters of interest. Inference on the original scale of the response makes the results easy for interpretation. Prior information about parameters can be incorporated in the model. The advantages of the new approach make it an attractive alternative to the standard statistical methods for antimicrobial assay analysis.

Small Area Estimation of Diabetes Prevalence

Haitao Chu (1), Lance A. Waller (1), and Deborah Rolka (2)
Department of Biostatistics, Rollins School of Public Health, Emory University (1)
Division of Diabetes Translation, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention (2)

OBJECTIVES: This presentation compares various small area estimators of county-specific prevalence of diagnosed diabetes, based on statewide Behavioral Risk Factor Surveillance System (BRFSS) surveys.

METHODS: We compare design-based "synthetic" estimation to model-based estimation based on Bayesian hierarchical logistic regression models with and without random effects, adjusted for age, gender, and race. Linking census data to the BRFSS data in a model-based format allows prediction in counties unsampled in the original BRFSS. Incorporating random effects in hierarchical logistic regression models results in "borrowed strength" across counties within the same state, and across local neighboring counties using spatial autocorrelation. Relating fixed effects to person-specific covariates and random effects to county-specific covariates in a multi-level model allows spatial similarity in prevalence rates between neighboring counties after adjusting person- specific and county-specific covariate effects. We consider four different structures for the prior distributions of the random effects, namely:

1) Exchangeable covariate effects and random intercepts with a CAR (conditional autoregressive) structure;
2) CAR covariate effects and exchangeable random intercepts;
3) CAR covariate effects and random intercepts; and
4) CAR covariate effects and a convolution prior for random intercepts.

RESULTS: We compare results for several states in the Southeastern United States, including Georgia (159 counties, 7 unsampled). We illustrate increased precision in the model-based estimates with little apparent bias, when compared to design-based synthetic estimates.

CONCLUSION: Compared to design-based synthetic estimates, our multi-level model-based estimates have slight changes in mean (increase in bias), but greatly reduced variation. The Bayesian hierarchical estimates result in larger variation than the logistic regression based estimates, reflecting uncertainty in parameters (nuisance parameters in the logistic model, hyperparameters in the Bayesian formulation) ignored by the logistic model, yet still gain precision over the design-based estimates.

Comparison of Hierarchical and Conventional Models of the Relation of Individual Life Events to Preterm Delivery

Nedra Whitehead and the PRAMS Working Group
Division of Reproductive Health, National Center for Chronic Disease Control and Health Promotion, Centers for Disease Control and Prevention

OBJECTIVES: This study compares estimation of the relation between preterm delivery and each of 18 stressful life events using conventional maximum likelihood analysis, classical Bayesian analysis and two hierarchical modeling approaches, empirical-Bayes and semi-Bayes modeling. Empirical-Bayes methods estimate the prior mean and variance of the parameters from the study data and other external information. Semi-Bayes methods estimate the prior mean from the study data but allow the variance to be determined by the researcher. Bayesian methods were used to incorporate existing knowledge about these relationships into the estimates, to increase the precision of the estimated parameters for the multiple exposure variables, and to alleviate multiple inference concerns.

METHODS: Study data were from the Pregnancy Risk Assessment Monitoring System (PRAMS), an ongoing study which surveys a state population-based sample of women who delivered a live born infant 2-6 months previously. The analysis included 10,274 multiparous women with singleton deliveries in 1990 though 1993. Maximum likelihood estimates were obtained using conventional logistic regression. For the classical Bayesian analysis, the prior distributions were determined from an extensive literature review of factors related to preterm delivery. For the empirical-Bayes analyses, the prior mean and prior covariance matrix was estimated from the data using methods described by Witte. These methods were also used to estimate the prior means for the semi-Bayes analysis, but the prior variance was determined from the literature. Posterior distributions were calculated as weighted averages (based on the variances) of the prior distribution and the maximum likelihood estimates.

RESULTS: None of the life events examined were associated with an increased risk of preterm delivery. With one exception, the conventional odds ratios were centered around 1 (range: 0.61 to 1.20) and the 95% confidence intervals included 1. The death of a woman's husband or partner in the year before delivery was associated with a reduced risk of preterm delivery (OR: 0.34, 95% C.I.: 0.19, 0.59). There was little difference between the conventional and hierarchical estimates. For rare events, the odds ratios generally became closer to 1 and the Bayes posterior intervals are narrower than the confidence intervals.

CONCLUSION: The stressful life events studied here are not associated with an increased risk of preterm delivery. The association between being widowed during pregnancy and a reduced risk of preterm delivery requires further investigation to determine its significance. In this very large data set, there was little difference between conventional and hierarchical modeling, or between the three different Bayesian techniques.

Contributed Session 1C
Tuesday, January 23, 2001
11:00 – 12:30

Cancer Incidence and Radio Frequency Radiation Emitting from TV Towers: Investigating an Excess Risk Using Isotonic Regression Models.

Max Bulsara (1), Tony Morton-Jones (2), and Nick DeKlerk (3)
Department of Public Health, University of Western Australia (1)
Department of Mathematics and Statistics, University of Lancaster (2)
Department of Biostatistics and Genetic Epidemiology, Institute of Child Health Research, University of Western Australia (3)

OBJECTIVES: There is considerable public concern about the possible health effects of exposure to radio frequency radiation. The aim of this study was to examine if there is an excess risk of cancer associated with exposure to non-ionizing radiation from television transmitters.

METHODS: There are 4 television VHF transmitters operating within metropolitan Perth in Western Australia, serving the 4 free-to-air networks. Incidence data for brain cancer & leukemia from 1982-1996 was obtained from Western Australian Cancer Registry, with the individual address of each case at the time of diagnosis geocoded and aggregated into Census Collectors District (CD). Population data was obtained from the Australian Bureau of Statistics, aggregated into CD's by age group and gender. Socio-economic Indexes for Areas (SEIFA) was calculated for each CD and treated as a confounding variable as well as age and gender. In the analysis, distance from TV towers was used as a surrogate measure of exposure. Actual exposure data was obtained from discrete, randomly selected locations using spectrum analyzer measured power intensity specific to each transmitters. Recent developments in isotonic regression which allow covariate adjustment of, and confidence intervals for, the risk estimate as a monotonic function of explanatory variable(s) under a Poisson setting is to be used to model each CD specific risk estimate, with the risk estimate function tested against the null hypothesis of constant risk with respect to distance from the tower or power intensity.

RESULTS: At the present time, we have completed the data collection and handling, and the requisite methodological developments are in place to carry out the analysis, the results of which will be presented and discussed at the Symposium.

CONCLUSION: To be presented at the Symposium.

Probabilistic Model to Estimate Male-to-female Sexual Transmission of HIV-1

Hrishikesh Chakraborty (1), Pranab K. Sen (2), Ronald W Helms (2), and Myron S. Cohen (3)
Department of Biostatistics, Rollins School of Public Health, Emory University (1)
Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill (2)
School of Medicine, University of North Carolina at Chapel Hill (3)

OBJECTIVES: Epidemiological and mathematical models have been developed to estimate the likelihood of HIV-1 transmission during a single episode of sexual intercourse. Most published estimates assumed constant infectivity within and between couples. We developed a probabilistic model to estimate the male-to-female penile-vaginal per-sexual-act HIV-1 transmission probability as a function of changes in seminal viral load and receptor cell (CD4+CCR5) numbers.

METHODS: We mainly used conditional and unconditional probability theory to develop the model. Pearsonian type-1 and Beta distribution was also used to develop the model. We also used bootstrap re-sampling method to account for repeated observations and successive approximation method to estimate model parameters.

RESULTS: We evaluated data sets from three different centers (n=88, 165, and 100) for seminal plasma RNA concentrations and one dataset (n=28) for receptor cell counts. The Beta distribution parameter estimates for seminal plasma HIV-1 RNA NSI count in one ejaculate for CD4 < 200 group were a11=0.385, b11=5.646 and for CD4 > 200 group were a12=0.242, b12=1.428. The Beta distribution parameter estimates for receptor cell counts/mm2 were a21=0.769, b21=1.143. The model predicts that the per-contact HIV-1 transmission probability ranges from 0.0001 to 0.0003 when the seminal viral load is 1,000 copies per ejaculate and the receptor cell count ranges between the 25th to 75th percentile. The model demonstrates a sharp increase in transmission probability as seminal viral load and/or receptor cells count increases. For 100,000 copies of HIV-1 RNA in one ejaculate, the transmission probability ranges from 0.0039 to 0.0096.

CONCLUSION: Estimates of the efficiency of transmission of HIV-1 have been derived from epidemiological studies and mathematical models. However, the transmission probabilities presented were so low that it becomes difficult to understand the magnitude of the HIV-1 pandemic, especially in developing countries. The results suggest a per-contact rate of sexual transmission of HIV-1 that better explains the magnitude of the epidemic than older epidemiological models. Our model can be used to examine the biological basis for accelerated spread of HIV-1 in some developing countries, and the effects of different prevention strategies that influence viral burden, viral phenotype, and expression of receptor cells. Such an approach can be expected to lead to the most rational and efficient use of limited resources.

An Investigation of Error Sources and Their Impact in Estimating the Time to the Most Recent Ancestor of Spatially and Temporally Distributed HIV Sequences

Tom L. Burr (1), James R. Gattiker (1), and Philip J. Gerrish (2)
Safeguards Systems Group, Nonproliferation and International Security Division, Los Alamos National Laboratory (1)
Theoretical Biology Group, Theoretical Division, Los Alamos National Laboratory (2)

OBJECTIVES: This study is an investigation of some major error sources and their impact in estimating the time to the most recent common ancestor (MRCA) of spatially and temporally distributed human immunodeficiency virus (HIV) sequences.

METHODS: We use coalescent theory to simulate an HIV epidemic under a range of assumptions about how the epidemic is progressing. We assume that all present day HIV-1 (subtype M) sequences arose from one HIV-1 ancestor, following one simian (chimpanzee) to human transmission. This allows us to simulate sample genealogies with known time to the MRCA. We then apply a range of baseline ("known") evolutionary models to generate sequence data. We next estimate or assume one of several misspecified models and use the chosen model to estimate the time to the MRCA. The extent and type of model misspecification determines the magnitude of our error sources that could include any of: neglected heterogeneity in substitution rates across lineages and DNA sites, uncertainty in HIV isolation times, uncertain magnitude and type of population subdivision, uncertain impacts of host/viral transmission dynamics, and unavoidable model estimation errors.

RESULTS: Our simulation results suggest that confidence intervals will rarely have the nominal coverage probability for the true time to the MRCA. The reason is that neglected effects lead to systematic errors that are unaccounted for in most analyses, resulting in optimistically narrow confidence intervals. Using real HIV sequences having approximately known isolation times and locations, we then present possible confidence intervals for each set of assumptions.

CONCLUSION: In general, we cannot be certain how much to broaden a stated confidence interval for the time to the MRCA. However, we describe the impact of candidate error sources on confidence interval width. We also discuss which error sources have the most impact on confidence interval width for the time to the MRCA of HIV.

Treatment Patterns and Health Outcomes: Analyzing Differences in Medical Practices

Jeffrey J. Geppert (1), Douglas O. Staiger (1,2), and Mark B. McClellan (1,3)
Health Care Economics Program, National Bureau of Economic Research (1)
Department of Economics, Dartmouth College (2)
Departments of Economics and Medicine, Stanford University (3)

OBJECTIVES: The randomized controlled trial is the "gold standard" in estimating effect of medical treatments on patient outcomes, but RCTs have limitations for evaluating effectiveness in a wide range of treatments and populations. The two main approaches to using observational data for estimating treatment effects include adjustment for observable patient differences and instrumental variables. The limitations of adjustment include the availability of satisfactory risk adjusters, the cost of data collection, the effect of other treatments on outcomes, and the direct impact of survival on treatment. The limitations of instrumental variables include the difficulty of finding valid instruments and the application when many treatments must be considered.

METHODS: We develop a new approach that evaluates the effects of different practice patterns, using between-hospital variation in treatment intensity as instrument to account for small hospital samples (weak instruments). We estimate provider fixed effects for each outcome and treatment measure using multi-variate signal extraction methods, related to empirical Bayes methods but less computationally intensive. Factor models summarize multidimensional differences in treatment across hospitals, which allows estimation of the extent to which identifiable practice differences are associated with differences in hospital outcomes, and the impact of these treatment differences. We check validity of our results using between-zip code variation in treatment intensity to estimate practice pattern effects (using zip code intensity as instrument).

RESULTS: Six factors summarize systematic differences in hospital practice patterns for AMI, which provides summary of differences in practice patterns and their relationship to patient outcomes. The most important factor for describing hospital mortality is closely related to aspirin and beta blocker use, and use of some other drug treatments also shown to reduce heart attack mortality. Other significant factors related primarily to invasive procedure use (angioplasty, bypass); all associated with lower mortality, as expected. Similar results for zip-level analysis of practice patterns, which is not biased by hospital selection. Given other treatments, no clear association between thrombolytic & ace inhibitor use and better outcomes.

CONCLUSION: We provide a robust, general method for estimating treatment effects with observational data. We use variation in treatment intensity across hospitals and zip codes which substantially eliminates bias arising from treatment selection being based on unmeasured patient factors. Practice pattern analysis can summarize treatment differences across hospitals, and determine which practices are most strongly associated with better outcomes.

Contributed Session 2A
Tuesday, January 23, 2001
2:00 – 3:30

Comparison of Hybrid Tree Predictor/Logistic Regression When Both Dependent and Independent Variables Are Dichotomous

Lawrence E. Barker and Cedric Brown
Data Management Division, National Immunization Program, Centers for Disease Control and Prevention

OBJECTIVES: Tree models, such as those derived using CART™ or the tree algorithm of SAS Enterprise Miner™ and models derived using binary logistic regression both can predict dichotomous variables. Tree models very effectively model local effects, but are less effective with global effects. The situation is typically reversed for logistic regression. Hybrid tree/logistic regression methods that effectively model both global and local effects have been proposed. For example, one might: (1) use dummy variables for terminal nodes in a tree model as additional independent variables for logistic regression or (2) use a linear combination of independent variables, obtained via logistic regression, as an additional independent variable for a tree model. The relative merits of methods (1) and (2) have not been studied. We compare these methods and provide recommendations concerning when each should be used.

METHODS: Methods are compared via simulation, where a variety of dependency structures between dependent and independent variables are considered. Logistic regression outcomes are converted to dichotomous predictions using an algorithm comparable to that used in the tree model. The tree algorithm of SAS Enterprise Miner™ is used. For simplicity, only dichotomous independent variables are considered. Comparison is primarily through sensitivity/specificity of models derived using methods (1) and (2), although ease of interpretation is considered.

RESULTS: Tables of approximations to sensitivity and specificity of models derived using methods (1) and (2) are given for a variety of dependency structures. Models derived using method (1) are more easily interpreted than those derived using method (2). Method (2) is more severely impacted by missing values than method (1), since the linear combination of variables obtained from logistic regression depends on all variables, and terminal node membership frequently depends on only some variables.

CONCLUSION: Method (1) should be used if the data contain many missing values. If few values are missing, either method might be considered. Sensitivity/specificity of models obtained using methods (1) and (2) are roughly comparable under a variety of circumstances, although differences exist. Choice should be guided by the relative importance, in a particular application, of sensitivity/specificity and ease of interpretation.

Characterizing the Functional Form of a Neural Network Model

Douglas Landsittel (1), Harshinder Singh (2), and Vincent C. Arena (3)
Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention (1)
Department of Statistics, West Virginia University (2)
Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh (3)

OBJECTIVES: Traditionally, neural networks have been considered black box techniques which are only useful for classification purposes. To expand their utility in biostatistical applications, numerical approximations and Monte Carlo simulations were implemented to represent neural networks as interpretable polynomial functions, thus elucidating the underlying structure of the corresponding data set.

METHODS: Network structures considered in this analysis were limited to a single hidden layer, variable numbers of binary inputs and hidden units, and a single binary outcome. Through Taylor series expansions, the neural networks can be represented as a function of the inputs, with an undetermined number of interactions. Since the optimal network weights are determined using a criteria equivalent to maximum likelihood, likelihood ratios between a neural network and the intercept model follows a chi-square distribution with the degrees of freedom equal to the number of independent parameters. Monte Carlo simulations were conducted to calculate the distribution of the statistic under the null hypothesis of no association. The resulting simulated degrees of freedom and the Taylor series expansion, were considered to define the functional form of the fitted model.

RESULTS: For binary input variables, neural networks fit all possible main effects and interaction terms given a sufficient number of hidden units. The resulting number of independent terms can be generally defined by the minimum of two quantities: 1) the total number of main effects and interactions, and 2) the number of weights in the network. Simulations confirm that these results hold empirically and that the likelihood ratio follows a chi-square distribution asymptotically. Correlations between the coefficients of Taylor's expansion were calculated under different associations to elucidate the functional form of the model when the number of weights is not sufficient to estimate all terms. In such cases, adding an additional input to the model corresponds to H additional independent terms, where H is the number of hidden units in the network.

CONCLUSION: This study interprets the functional form of a neural network model through polynomial approximations of the network response function and simulation results. In addition to providing a framework for hypothesis testing of binary inputs, results lead to insight into the number of independent parameters fitted by the model and the polynomial functional form of the model. Such findings will contribute significantly to relevant medical applications by providing a statistical context for selecting and interpreting network architectures, which has previously been a relatively ad hoc process.

Comparing Two Goodness-of-split Measures for Classification Trees: a Simulation Based Study

Mousumi Banerjee (1) and Anuradha Roy (2)
Center for Healthcare Effectiveness Research, School of Medicine, Wayne State University (1)
Department of Mathematics and Statistics, Oakland University (2)

OBJECTIVES: Tree-based methods are adaptive nonparametric statistical procedures that are gaining popularity in medicine and the health sciences, as a result of increasing complexity of study designs and data structures. Two most commonly used softwares for performing tree-based analyses are S-plus and CART. These employ different impurity measures for node-splitting, namely the deviance and the Gini index in S-plus and CART respectively. This paper compares the performance of these two goodness-of-split measures in the context of a binary classification problem.

METHODS: An extensive simulation was conducted to compare deviance with the Gini index in terms of classification accuracy. For each simulation run, we initially generated trees that were fully grown. Next, optimal trees were generated from the fully-grown trees using cost-complexity pruning and ten fold cross-validation. The two measures were also compared in terms of the size of the resulting optimal trees, since comprehensibility of tree structure is related to tree size. The effect of the addition of independent noise variables on classification accuracy and tree size was also investigated for each measure.

RESULTS: The fully grown as well as optimal trees produced in S-plus and CART were different in architecture as well as tree size. Optimal trees using deviance as the splitting criterion are consistently bigger in size (roughly 2-3 times) than the optimal trees that are obtained using the Gini index. When the learning sample has roughly equal proportion of cases for each response category, the misclassification error rates of trees based on Gini index and deviance are sufficiently similar that their differences are statistically insignificant. However, in situations where one response category is disproportionately sparse (i.e. comprising $< 20\%$ of the learning sample), the misclassification error rates of trees based on the Gini index is 1.5-2.5 times higher than that of trees based on the deviance measure. This holds true for similar sized trees as well.

CONCLUSION: In practical applications, if small differences in misclassification rates are not important, the user may wish to select the Gini-based tree (because of its easy comprehensibility) when the learning sample has roughly equal proportion of cases for each response category. In applications where one response category is sparse, the deviance-based trees are superior to the Gini-based trees in classification accuracy.

Sequential, Spatial Stochastic Curtailment

Tim E. Aldrich (1), Guang Zhao (2), Jim Ferguson (2), and Murray B. Hudson (2)
Preventive Medicine, Department of Family and Preventive Medicine, University of South Carolina School of Medicine (1)
Public Health Statistics and Information Services, South Carolina Department of Health and Environmental Control (2)

OBJECTIVES: A balloonist has a high-level view of a community, but nonetheless must often descend to closer levels to discern specific details. Similarly in public health surveillance, analyses at multiple levels may be used with a progressive logic when disease risk signals are needed and resources for fieldwork and in-department data collection are not available. In such circumstances, simple accruing probability may be applied as a means for identifying settings to follow-back for most productive findings.

METHODS: We are preparing statistical solutions for such application using prior distributions to form a population-based probability-weighted geographic surface. This approach is analogous to the density equalized map projections performed to redesign surface boundaries using an underlying population distribution. However, this application relies on established disease occurrence patterns, e.g., risk factors, as well as population characteristics. The age-, race-, gender-, space-specific prior distributions provide a Bayesian 'expected' probability surface for mapping cases sequentially as they occur. One may select sentinel events for biologic value from suspected environmental hazards, or high-risk configurations for under-served groups (e.g., rural populations, children) to provide a hierarchical weighting to cases or to findings. Then, sequential solutions may be programmed with false positive protection and early recognition thresholds for recognition of emerging patterns. The test distribution is a negative binomial solution and takes a form analogous to the sequential methods applied with clinical trails decisions.

RESULTS: We have simulated these applications to illustrate these 'balloonist' applications of sequential observation for declining spatial variability with case distributions for asthma, diabetes, and cancer. Illustration of the progressive logic of the spatially declining randomness within a case-series will be described. Our follow-back studies of these findings revealed instances of bona fide disease risk, as well as system classification system errors, both results vindicate the surveillance application.

CONCLUSION: This strategy with public health surveillance can involve multiple-level analyses, e.g., statewide, regions within a state, counties, and for some metrics even ZIP codes. The 2000 census also offers a profound opportunity for census tract-based spatial analyses.

Contributed Session 2B
Tuesday, January 23, 2001
2:00 – 3:30

Evaluating Effective Degrees of Freedom in Longitudinal Data Sets

Michael Brimacombe
Department of Preventive Medicine and Community Health, New Jersey School of Medicine, University of Medicine and Dentistry of New Jersey

OBJECTIVES: In many longitudinal settings, especially in the early stages of experimentation and study, the use of excessive numbers of time points in the collection of data may artificially expand sample size and render p-values unstable (always significant). The use of deterministic, functional data models is examined as a means of obtaining lower bound estimates of effective degrees of freedom in the dataset. These are used to both obtain pseudo p-values for comparative purposes and also correct standard p-value calculations.

METHODS: For many physiological response measurements the use of chaotic, deterministic functions helps explain much of underlying response patterns. If such functions are assumed to generate the response in question, repeated measures ANOVA designs can be viewed as standard ANOVA models with a functional data response. In such a setting fewer assumptions regarding underlying correlation are required and functional data analytic tools can be employed to obtain average response within and between treatment for which ANOVA type testing can be developed. Also available are degrees of freedom. The issue of effective degrees of freedom can be examined in this setting as the assumed chaotic-functional data structure can be viewed as providing lower bounds on the number of independent sources of random variation present in the observed data.

RESULTS: In settings where there is clear oversampling due to excessive design time points, standard p-values may not be stable. Correcting the degrees of freedom available using the functional data model values is suggested. Examples are shown where significance in standard p-value calculations is clearly an artifact of excessive sampling. These calculations are corrected using degrees of freedom obtained using the functional data model. They are also compared to p-values obtained from the functional data model.

CONCLUSION: Restricted assumptions regarding independent sources of variation can help determine more appropriate degrees of freedom that can then be used to help adjust for excessive sampling, especially through time, that leads to always significant p-values.

Creating and Disseminating Bootstrap Weights: the Canadian NPHS Experience

Francois Brisebois, Guylaine Dubreuil, and Denise Hall
Household Survey Methods Division, Statistics Canada

OBJECTIVES: The Canadian National Population Health Survey (NPHS) is a longitudinal survey conducted by Statistics Canada which aims to provide data for analytic studies to help in understanding the determinants of health. For this purpose, NPHS collects data on the economic, social, demographic, occupational and environmental correlates of health. For many reasons, notably because of the complex nature of its multistage design, NPHS has recourse to the bootstrap sample reuse method to generate weights for variance estimation. The focus of the presentation is to relate the different aspects surrounding the use of bootstrap weights, from their creation to their dissemination.

METHODS: The bootstrap method is considered as a simple and practical way to allow users to calculate their own variances for survey estimates. Variance estimates are obtained by comparing estimates produced using weights generated from bootstrap subsamples. Bootstrap weights also have the advantage that they allow the incorporation into the variances the effects of adjustments for nonresponse, for interprovinicial movers, and for post-stratification by provincial age-sex groups. Bootstrap weights were initially supposed to be provided along with the NPHS public-use microdata files. However, mainly because of confidentiality reasons, this strategy had to be rethought.

RESULTS: Since the bootstrap weights could not be released with public-use files, an alternative had to be found to still provide external users with a way to estimate proper variances. The adopted alternative was to provide public-use file users with dummy bootstrap weights files. These files would have the same structure as the real bootstrap weights files, but these dummy bootstrap weights would contain perturbed data. Users can use these dummy files to prepare their variance estimation programs which, once duly tested, can be sent to Statistics Canada where they are submitted using the actual data and bootstrap weights files. This service is part of what is referred to as the Remote Access Services. The entire process is well documented and users have responded favourably. Major issues concerning the support and practicality of this service will be discussed during the presentation.

CONCLUSION: NPHS is currently undertaking its fourth collection cycle and its experience with bootstrap weights is very satisfactory. The presentation of many workshops across the country promoting this innovative approach, the continuous effort in improving our technical support, and the constant monitoring of the efficiency of the bootstrap weights are, among others, ongoing NPHS activities that will also be discussed during the presentation.

Follow-up Designs and Planned Missing Data in Repeated Measures of Longitudinal Epidemiologic Studies

Alka Indurkhya
Division of Biostatistics, Michigan State University

OBJECTIVES: Most longitudinal epidemiologic studies find it necessary to continually update their follow-up study designs due to either budget limitations such as available funding or logistic considerations. The goal here is to provide guidelines for choosing follow-up study designs with planned missing data for repeated measures in a multi-stage sampling framework.

METHODS: In multi-stage sampling, one specific variable or a set of variables, can be used to determine the probabilities of observing the next stage repeated measures. Due to either budgetary or logistic constraints, one may need to determine how many and which subjects to follow-up at a given stage of the sampling framework while still providing consistent and asymptotically unbiased estimators of primary interest to the long-term goals of the study. Thus, the follow-up design choices may involve planning missing data (ignorably or nonignorably) either with or without stratification. Data can be missing ignorably if they are missing completely at random or missing at random. Otherwise data is said to be missing nonignorably. The choice of follow-up designs under these planned missing data mechanisms are evaluated by performing power calculations (using a Markov Chain Monte Carlo approach) for the main scientific hypotheses of the study.

RESULTS: Not all follow-up designs lead to identifiable parameters. In order to select designs that have identifiable estimators for all parameters of interest, the problem of design selection is reformulated as a problem in pattern mixture models.
In particular, under the mechanism of nonignorably missing data it is found that the key study parameters are identifiable only if the missing mechanism for currently observed data does not depend on the determinants of planned nonignorably missing data at future follow-up of subjects. This restricts the number of possible follow-up designs with nonignorably planned missing data.
The methodology is illustrated by presenting future follow-up study design choices using data from an ongoing longitudinal epidemiologic study that examines the long-term benefits of early classroom based interventions on outcomes such as graduation, reduced juvenile delinquency, and substance abuse.

CONCLUSION: This work provides an analytic strategy for selecting follow-up study designs midway into multi-stage longitudinal studies when missing data is planned as part of future follow-up measures and yet provide consistent and asymptotically unbiased estimators for study parameters of interest.

Illness-death Stochastic Model in the Analysis of Longitudinal Dementia Data

Jaroslaw Harezlak (1), Sujuan Gao (1), and Siu L. Hui (1,2)
Division of Biostatistics, Indiana University School of Medicine (1)
Regenstrief Institute for Health Care (2)

OBJECTIVES: Longitudinal epidemiological studies on dementia usually use two-phase sampling designs to estimate disease incidence and risk factors associated with it. Current statistical methods are not adequate to deal with all issues inherent in those studies, namely, longitudinal outcomes, missing data due to death of elderly subjects and complex sampling designs. Invalid assumptions made to analyze the data may lead to biased results. We propose a method dealing with all aforementioned issues in order to draw valid conclusions.

METHODS: We propose a stochastic model approach to simultaneously estimate disease incidence and mortality rates. We set up a Markov chain model consisting of three states: healthy, diseased and dead. We assume various hazard functions and estimate the transition probabilities using maximum likelihood approach. Constant age-specific hazard rates and hazard rates dependent on covariates are considered. Further, we extend our model to incorporate complex sampling design used in dementia studies.

RESULTS: We conducted a simulation study in order to assess the performance of the proposed method. We obtained substantial improvements in the estimation of the constant hazard rates. Average relative bias is within 0.5% of the true value, compared with 3% to 7% for the naive estimator. Also, we obtained good estimates of the covariates' coefficients. Furthermore, we will apply our model to data from a community-based dementia study.

CONCLUSION: The proposed stochastic model approach performs well in simulation studies. It effectively corrects the bias in naive estimators that are ignoring deceased subjects.

Contributed Session 2C
Tuesday, January 23, 2001
2:00 – 3:30

Toward an Operational Definition of "Personally Identifiable Information"

Steven M. Banks and John A. Pandiani
The Bristol Observatory

OBJECTIVES: To provide an operational definition of "personally identifiable information" that will allow health researchers and data archive administrators to make reasonable decisions about access to data. This operational definition should be sensitive to both the personal privacy of individuals and the need of program administrators, health care providers, payers, and the public to obtain information on the performance of hearth care providers and systems of care.

METHODS: Statistical procedures derived from probability theory are used to determine the population size necessary to achieve specified levels of identifiability. These statistical procedures are applied to the problem of determining the size of a community in which the combination of date of birth and gender may be considered personally identifying information.
The determination of the probability that more than one person in the population will share a birthday (month and day) is provided by the solution to the classic birthday problem. Similar logic provides the probabilities related to the sharing date of birth (month, day, and year) and gender.
The probability that a specified person will share his or her date of birth and gender with another person in a population is calculated as the average probability of sharing and as the minimum probability of sharing.

RESULTS: Date of birth and gender tend to be personally identifiable in communities of less than 2,600 people. When populations reach 72,000, false identification is more likely than correct identification of individual people.

CONCLUSION: Data sets that include the date of birth and the gender of individuals in conjunction with county of residence should rarely be considered personally identifying. The statistical procedures described in this presentation can be applied to any set of personal characteristics when the distribution of these characteristics in the population is known. This information provides a rational basis for making decisions about access to data sets.

The ability to know the degree to which date of birth and gender should be considered to be personally identifying is particularly important in light of the ability of Probabilistic Population Estimation to measure critical treatment outcomes. Probabilistic Population Estimation is a statistical technique for measuring the number of people represented in both of two anonymous data sets (e.g. a treatment data set and a mortality data set) based on the distribution of date of birth and gender in the data sets (Banks and Pandiani, Statistics in Medicine, in press). This methodology is particularly important in this era of heightened concern about the protection of personal privacy (Pandiani and Banks, Journal of Behavioral Health Services and Research 25:4, 456-463).

A Web-based Database System Design for Impact Evaluability Assessment of HIV Prevention Efforts and Policies at the Municipal, State, and National Levels

Xiaohong Mao Davis, Choi K. Wan, Timothy A. Akers, and Huey-tysh Chen
Division of HIV/AIDS Prevention, National Center for HIV, STD, and TB Prevention, Centers for Disease Control and Prevention

OBJECTIVES: Identify relevant data sources that can be integrated into a relational database for conducting impact evaluability assessment of HIV prevention efforts and policies at the municipal, state and national levels. Integrate multiple types and levels of HIV indicator data, policy data, risk behavior data, demographic data, community planning data and CDC financial data into the relational database. Support multiple software analytic tools such as AUTOBOX, SAS and SPSS for performing quasi-experimental analysis, such as interrupted time-series analysis.

METHODS: A web-based relational database that can support a wide variety of users and stakeholders will be developed using MS SQL Server (v7.0). Relevant data will be collected from multiple existing data sources within CDC and other federal agencies. Variable attributes and information will be documented and stored within database in the form of metadata. Same variables using different measurement scales due to different data sources will be standardized following the CDC Common Data Elements and other commonly used conventions. New data can be dynamically added to the system. Database management and Web management will be handled by two different small servers as they are more cost-efficient than one big server that attempts to perform both tasks.

RESULTS: This data system, in response to the need for an integrated set of data to explore the relationship between CDC's funding of HIV prevention efforts, the prevention activities that are implemented, and multiple software analytic tools can be used to extract the data for modeling. The main data set is stored in simple flat tables and the relationships among data are represented by storing additional data columns. Multiple types of variables will be accommodated in the same tables and additional variables can be integrated into the system with special permission. Internet-based design will facilitate dissemination to a wide variety of users and stakeholders.

CONCLUSION: This is the first effort within CDC to systematically collect HIV prevention data from a variety of data sources such as behavioral and surveillance data systems. Data identified in the five domains (biological, behavioral, interventions, policy and fiscal) related to HIV prevention provides a valuable resource for conducting impact evaluability assessment on past and current HIV prevention efforts over time. For example, evaluability assessments can be conducted on major HIV prevention efforts to determine which areas are feasible to carry out impact evaluation and which are not. Practical strategies can then be explored for conducting and enhancing impact evaluability for those areas where it is not feasible to carry out impact evaluation. Building such a comprehensive database promotes and strengthens collaboration among different units within CDC and other federal agencies.

Cancer in the Disabled Medicaid Population in Oregon: Data Collection and Management

Jodi Lapidus (1), Donald Austin (1), Hank Bersani Jr. (2), Adam Evans (3), and John Hough (4)
Department of Public Health and Preventive Medicine, Oregon Health Sciences University (1)
Division of Special Education, College of Education, Western Oregon University (2)
Department of Public Health and Preventive Medicine, Oregon Health Sciences University (3)
Division of Birth Defects and Developmental Disabilities, National Center for Environmental Health, Centers for Disease Control and Prevention (4)

OBJECTIVES: This project is collaboration between the Oregon Health Sciences University (OHSU), Oregon Medical Assistance Programs (OMAP), Oregon State Cancer Registry (OSCaR) and Oregon Senior and Disabled Services Division (SDSD). The goal is to estimate the risk of various cancers in the population of Oregon Medicaid clients with disabilities. We hypothesized that, in this population, certain cancers would occur more frequently, and certain cancers will be diagnosed at later stages than in non-disabled groups.

METHODS: Several steps were required to collect, manage, link, store and analyze large data files from these different state agencies. Medicaid eligibility data files from 1996-1998 identified our population of interest. Since name and address were needed to link to the cancer registry, we merged Medicaid-eligible clients to OMAP name and address databases and created a "Medicaid population file" to be linked to the cancer registry. All files housed within OMAP were merged using their unique Medicaid identification number. OSCaR does not store Medicaid identification number, so substantial review was required to prepare the Medicaid population file for the cancer registry linkage. We then used probabilistic matching techniques, based on social security number, name, gender, date of birth and address, to link the cleaned Medicaid population file to OSCaR. To identify the disabled clients on Medicaid, we used OMAP program eligibility codes, which identify the basis for eligibility, supplemented by data from a linkage to the SDSD client assessment database, which identifies functional status and special services needed for disabled clients. Merged data files were further reviewed to ensure that the appropriate clients would be used in the numerator and denominators of the rate calculations. Finally, reduced-size analysis files were created containing variables pertaining to hypotheses of interest, and any client-identified information was encrypted and moved to backup tapes for confidentiality purposes.

RESULTS: We identified over 1 million unique Medicaid clients in Oregon between 1996 and 1998. Approximately 4000 matches occurred between the Medicaid population file and the cancer registry. Functional limitation information on 10% of the Medicaid population was obtained via SDSD database. Analysis of cancer risk is in progress.

CONCLUSION: This successful collaboration demonstrates how a state cancer registry and other existing administrative databases can be used to identify cancer risk in a population of interest.

Use of Replication Methods in the National Immunization Survey to account for Nonsampling Weight Adjustments

Philip J. Smith (1), K.P. Srinath (2), Michael P. Battaglia (2), David C. Hoaglin (2), Barry I. Graubard (3), Lawrence Barker (1), Martin Frankel (2), and Meena Khare (4)
Data Management Division, National Immunization Program, Centers for Disease Control and Prevention (1)
Abt Associates Inc. (2)
National Cancer Institute, National Institutes of Health (3)
Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention (4)

OBJECTIVES: Large-scale national surveys routinely involve adjustments of sampling weights. These nonsampling adjustments include weight trimming, compensation for unit nonresponse, poststratification, and adjustment for noncoverage of nontelephone households (if the survey is a phone survey). This paper aims to provide a better understanding of the effect of such adjustments on standard errors of survey estimates. In particular, different methods of estimating standard errors may produce substantially different results.

METHODS: To account for nonsampling weight adjustments that are made in the National Immunization Survey (NIS), we propose the use of a computationally intensive method known as replication methods.

RESULTS: Results from our case study suggest that failure to account for nonsampling weight adjustments may produce biased estimates of standard errors. This bias may in turn distort the coverage levels of confidence intervals and the significance levels of statistical tests.

CONCLUSION: These results are congruent with those given in related research by Rubin and Schenker (1986, 1987), who investigated the failure to properly account for statistical methods that adjust for other nonsampling errors such as imputation for item nonresponse.

Contributed Session 3A
Wednesday, January 24, 2001
11:00 – 12:30

Using Large Datasets from Multiple Sources:
Achieving Flexibility, Performance, and Data Quality, or: Data Encounters of the Third Normal Kind

Dana Abouelnasr (1), Mary E. Hewitt (2), Matthew Lederer (2), Steve Martin (2), Ronald Parker (1), and Uma Shanmugan (2)
Office of Program Operations and Management, Agency for Toxic Substances and Disease Registry (1)
Electronic Data Systems (2)

OBJECTIVES: To present our experience with assimilating environmental datasets from different sources and of varying formats, quality, and content. To discuss process techniques for dealing with the accumulated large volume of data.

METHODS: Environmental health professionals must often address many environmental datasets from many sources. Datasets are frequently incomplete, scantily documented, and lack key fields. Additionally, formats for environmental data are not standardized, with each dataset possessing unique characteristics. Thus, the data must be standardized before it can be analyzed.

In the search for data quality and integrity, we have established uniform data management procedures, ranging from administrative concerns, such as compiling metadata for each dataset, to formatting and standardizing techniques. The procedures also include checks for a plethora of potential problems, similar to ones encountered in past datasets. The checklist is constantly updated, ensuring that the data management process improves with age. We will discuss our approach to developing these data management procedures.

Once large datasets have been assimilated into a single database, response times can suffer, and test the user's patience. To obtain timely responses, the database designer must occasionally break the basic rules of database design and normalization. When should an item of data be stored in more than one table in the database? When should aggregate data be stored, versus calculated at run-time? When and how many indexes should be created? We have developed design techniques and used basic database features, such as indexes, triggers, and stored procedures, to improve performance dramatically.

RESULTS: The established data management procedures have made the process more efficient and less susceptible to error. The procedures also ensure that our experiences become institutionalized. Once data are processed, they are consistent across our database and with other agency databases.

Because the resulting data may be massive, performance issues must be optimized by using creative design techniques and capitalizing on database performance features.

CONCLUSION: Established data management procedures and improved performance allow users to evaluate massive datasets with confidence and good response times.

An Environmental Information Management System Integrating Textual, Graphical, and Geographical Interfaces

Dana Abouelnasr (1), Ronald Parker (1), Jason Curtis (2), Mary E. Hewitt (2), and Uma Shanmugam (2)
Office of Program Operations and Management, Agency for Toxic Substances and Disease Registry (1)
Electronic Data Systems (2)

OBJECTIVES: Environmental health analyses at federal sites are often based upon large amounts of sampling data. The Federal Facilities Information Management System (FFIMS) was developed to manage these data.

METHODS: The FFIMS architecture integrates several components, all remotely accessible through a Local Area Network. Environmental and toxicological data reside on a database server, while map features, such as TIGER files, reside in a data warehouse central to the agency. A mapserver provides the mapping capabilities.

A user interface assists in browsing the data, with features such as histograms of frequency, time-series, and substance-components. Users can screen contaminant concentration data against standard values, build suspected or proven pathways of exposure, estimate doses using a variety of statistics, and compare those doses to information collected from toxicological studies.

Users specify map characteristics in a menu-based form. Then, the geographic interface lets them zoom in or out, pan, turn the various geographic features on or off, and graphically delineate an area for further analysis. The close integration of the components allows the mapserver to communicate information about that area (such as demographic data and sampling information) to the database server. Subsequent analysis outside of, and beyond the capacity of the geographic software, can characterize the contamination within each area.

RESULTS: FFIMS is an expanding set of useful tools for evaluating large amounts of data. It has been used successfully at several sites.

CONCLUSION: FFIMS development will continue. We are currently developing a stand-alone version to use in the field for data entry and analysis of smaller datasets. We also plan to develop an internet version to provide FFIMS access to our partners.

Venue-based Sampling of Young Men Who Have Sex With Men (YMSM): A Practical Approach to Collecting Representative Data on a Hard to Reach Population.

Lillian S. Lin (1), Farzana B. Muhib (2), Melissa Cribbin (3), Ann Stueve (4), and Carolyn Guenther-Grey (1)
Division of HIV/AIDS Prevention, National Center for HIV, STD, and TB Prevention, Centers for Disease Control and Prevention (1)
TRW (2)
CDC Information Systems Support Services (CISSS) (3)
Columbia University (4)

OBJECTIVES: We survey a cross-sectional, representative sample of YMSM in each of thirteen study communities over four summers as part of the evaluation of CITY, a CDC-funded, randomized trial of a community-level intervention to promote safer sex practices in YMSM.

METHODS: Effective access by project staff to YMSM was assured through a two-stage time-place sampling protocol. Before and during each survey wave in a community, a dynamic sampling frame was constructed of venues, venue-specific days and times within days (VDTs) when YMSM are likely to attend. Venues are identified through ethnographic methods. VDTs likely to produce adequate numbers of eligible respondents were determined through enumerations of venue attendees. VDTs were also categorized as "large" or "small" attendance, with cut-points varying by community. Each month, survey enrollment and new enumeration data were used to update the sampling frame. The first stage sample is a random selection of VDTs stratified by size. The selected VDTs are used to construct a 'survey event' calendar. In addition, up to three scheduled once-a-summer events can replace "large" VDTs each month. The second stage sample is collected during the survey events; interviewers visit the chosen venue and follow a systematic protocol to enroll eligible YMSM who visit during the time period. An Access® database was designed to track enumeration data for each VDT, to maintain the sampling frame, to conduct the first stage random selection of VDTs, and to store survey enrollment data. This information was electronically transmitted in a standardized fashion to the coordinating center by survey staff using programmed report generation functions.

RESULTS: Most surveys were conducted at gay-oriented (68%) or non-gay oriented (13%) commercial establishments. The remainder were collected at recurring (e.g. public sex environments) or special events (e.g. gay pride). Approximately 250 young men will have been surveyed in each community each year. The first wave resulted in 2621 usable completed interviews (87% of men who were eligible for interview). The second wave will be completed as of August 31, 2000. Enrollment success depends on the ability of survey staff to intercept venue attendees. The reporting system, including the VDT tracking program, resulted in timely reporting, accurate sampling frames, and frequent feedback to increase protocol adherence.

CONCLUSION: This is the only repeated cross-sectional, large-scale, community-based, representative survey of YMSM. The venue-based sampling protocol allows researchers to detect and accommodate differences in communities and changes over time and is easy to replicate with rigor.

A Record Linkage Technique for the Reconstruction of Individual Histories in Large Databases

Lara Lusa (1), Dario Gregori (2), and Annibale Biggeri (1)
Dipartimento di Statistica G.Parenti, University of Firenze (1)
University of Trieste (2)

OBJECTIVE: Record linkage techniques that are normally applied to merge databases or to detect duplicate records within them are not straightforwardly applicable to problems where an unknown number of records belong to the same subject, possibly containing information that are time varying. This situation canwell arise when there is the need to reconstruct individual histories using information contained in public health databases. The problem becomes even more complicated if there is the need to use an external database in orderto retrieve all the necessary information to match the records, therefore introducing higher error rates in the matching mechanism and, consequently, obtaining less precise results.

METHOD: A recursive method for comparison of records while reconstructing the individual histories is proposed to allow for multiple record linkage. The method takes into account the uncertainty introduced by the multiple linkage at each stage, allowing for more than one path for each subject to be followed, and eventually assigning to the reconstructed individual histories a score that measures the agreement between records. The mechanism borrows strength from the previously undertaken steps using a weighted information from the linkages already made to find the next best match, formally reducing to a one-by-one comparison of records. The obtained "agreement scores" for the individual histories can than be used to take into account the uncertainty arising from the matching mechanism in the statistical modeling of data and the more likely paths introduced simultaneously in the model.

RESULTS: The proposed method is applied to the database of the Italian Pacemaker Registry to reconstruct the individual histories of patients. Data of implantation and replacement of pacemakers are organized in two different databases; the same patient can have more than one implantation during his or her lifetime. Data cover the pacemaker implantation surgeries performed in Italy from 1980 till 1999 and the database referred to implants has more than 200.000 records. The individual histories of patients are reconstructed and the duration of the implant is modeled using a Cox's model where the uncertainty deriving from the record linkage mechanism is introduced in the frailty, using scores obtained through the process of record linkage for more than a single path.

CONCLUSION: The proposed method for record linkage allows for the reconstruction of individual histories in large databases and is particularly suitable when some statistical modeling of data is required, since the uncertainty related to the matching itself can be explicitly introduced in the model considering more than one possible path, associated with their "agreement scores."

Contributed Session 3B
Wednesday, January 24, 2001
11:00 – 12:30

Evaluation of Variance Components in an Mcf-7 Cell Culture Assay (E-Screen) Using Generalized Linear Mixed Models (GLMMs) and a Score Test

B. Rey de Castro and Donna S. Neuberg
Department of Environmental Health, Harvard School of Public Health

OBJECTIVES: The experimental design of biological assays is often based on a combination of well-founded theoretical and practical considerations aimed at achieving precise replicate observations, yet these designs routinely generate complex data structures that are not suitably analyzed by basic statistical techniques. The replicate observations may occur at multiple levels, constituting multiple variance components, and assay data may arise from an error distribution that may be difficult to specify prior to analysis.

METHODS: This paper utilized GLMMs to assess the statistical performance of an MCF-7 cell culture assay for estrogenicity known as E-SCREEN. This assay is replicated on 12-well cell culture plates, with each plate replicating the entire assayed dose range, so that there are three variance components: plate-to-plate variation; well-to-well variation, which is nested within the third variance component, the interaction between plate-to-plate variation and dose. GLMMs explicitly represent variance components through random effects parameters, and GLMMs admit a nearly universal range of error distributions, so that the optimal distribution for modeling the data may be identified. The performance of competing GLMM configurations was compared using information criteria. This study also applied a score test developed in 1997 by Xihong Lin (University of Michigan) to determine the statistical significance of variance components in GLMMs. SAS macros developed for this study to implement Lin's test are publicly available.

RESULTS: This statistical assessment of an MCF-7 cell culture assay found that a GLMM with a reciprocal link function and a gamma error distribution (est. COV = 2.9 percent) best represented the assay's dose-dependent effect. Lin's variance component score test found that each of the three variance components were statistically significant contributors to the overall variation in E-SCREEN data. These results were true for both 17b-estradiol, which induces maximal cell proliferation in E-SCREEN, and five weakly estrogenic polychlorinated biphenyls (PCBs 17, 49, 66, 74, and 128). Also, based on information criteria, the optimal gamma GLMM consistently out-performed equivalent naive normal and log-normal linear models, both with and without random effects terms.

CONCLUSION: This report demonstrates the efficacy of GLMMs and Lin's score test for analyzing complex data arising from an MCF-7 cell culture estrogenicity screening assay, especially for borderline estrogenic agents where a dose-dependent effect may be obscured if random effects are not accounted for. The techniques used in this paper may also be applied to other assays to characterize statistical performance and enhance experimental designs and quality assurance.

Nonparametric Functions and Random Effects: A Generalised Additive Mixed Models Approach

Marc Saez (1), Carmen Cadarso-Suarez (2), Adolfo Figueiras (2), and Javier Roca (2)
Research Group on Statistics, Applied Economics and Health (GRECS),Department of Economics, University of Girona, Spain (1)
University of Santiago de Compostela, Spain (2)

OBJECTIVES: We propose here the use of generalised additive mixed models (GAMMs) which are an additive extension of GLMMs. This extension uses additive nonparametric functions to model covariate effects while accounting for overdispersion and correlation by adding random effects to the additive predictors.

METHODS: We estimate the nonparametric functions by using smoothing splines. The extension of the generalised linear model incorporating random effects is called a generalised linear mixed model (GLMM). These models use a parametric mean function to model covariate effects. However, appropriate functional forms of the covariates may not be known in advance and the response variable may depend on the covariates in a complicated manner. It is hence of interest to allow more flexible functional dependence of the response on the explanatory variables.

In particular, we jointly estimate the smoothing parameters and the variance components by using marginal quasi-likelihood. This approach is an extension of restricted maximum likelihood (REML). Because numerical integration is often required by maximising the objective functions, double penalised quasi-likelihood (DPQL) is proposed to make approximate inference.

RESULTS: We illustrate the method by providing quantitative estimates of the short-term effects of air pollution on mortality in three Spanish cities, Barcelona, Valencia and Vigo for the period 1992-1992. Finally, we also evaluated the performance of the method through simulation and compared the approach with other alternatives. CONCLUSION: Our approach has proved to be simple, flexible and, at least, as efficient as known alternatives. We recommend the use of GAMMs in those occasions where appropriate functional forms are not known in advance.

CONCLUSION: Our approach has proved to be simple, flexible and, at least, as efficient as known alternatives. We recommend the use of GAMMs in those occasions where appropriate functional forms are not known in advance.

Estimating Correlation by Using General Linear Mixed Models: Correlation Between the Concentration of Blood and Semen HIV-1 RNA

Hrishikesh Chakraborty (1), Ronald W Helms (2), Pranab K. Sen (2), and Myron S. Cohen (3)
Department of Biostatistics, Rollin School of Public Health, Emory University (1)
Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill (2)
School of Medicine, University of North Carolina at Chapel Hill (3)

OBJECTIVES: Estimating correlation coefficient between two outcome variables is one of the most important aspects of epidemiological and clinical research. Simple Pearson's correlation coefficient method is usually employed when there are complete independent data points for both outcome variables. In practice, researchers normally deal with correlated observations in longitudinal setting with missing values where we are unable to use simple Pearson's correlation coefficient method. Our objective is to use GLMM techniques to estimate correlation coefficients when we have longitudinal measures with missing values.

METHODS: In this study, we used a random regression mixed model with unstructured covariance matrix to estimate correlation coefficients between log10 concentrations of HIV-1 RNA in blood and seminal plasma for different CD4 counts in a longitudinal setting with missing values. We also estimated correlation coefficients for patients during no antiretroviral therapy period and during an antiretroviral therapy period for different CD4 count. We compared the covariance and correlation structure for patients during no antiretroviral therapy period with during antiretroviral therapy period.

RESULTS: We used data sets from three different centers (n=88, 165, and 100) for blood and seminal plasma HIV-1 RNA concentrations. During the no-antiretroviral therapy period, 137 samples from 90 different patients and during the antiretroviral therapy period 513 samples from 148 patients were considered for analysis. We found no correlation during the no-antiretroviral therapy period between log10 blood and semen HIV-1 RNA concentration for CD4 counts ranges from 0 to 1,000. However, during the antiretroviral therapy period there was moderate correlation between log10 blood and semen HIV-1 RNA concentration for lower levels of CD4 counts (CD4 <200) and there was no correlation at higher CD4 counts (CD4>200).

CONCLUSION: Our findings confirm and extend earlier findings that factors influencing the infectivity of semen may differ from those that influence the infectivity of blood because male reproductive tract is a distinct immunological compartment. We concluded that there were significant correlation differences between the no antiretroviral therapy period and the antiretroviral therapy period. Our findings also suggested that there was a moderate correlation among the antiretroviral therapy groups at lower CD4 counts. Therefore, it is important to estimate and compare correlation estimates for different antiretroviral therapy patients to find appropriate therapies that reduce blood and seminal viral HIV-1 RNA at the same time because reducing seminal viral RNA has a direct impact on reduction of HIV-1 transmission.

Generalized Linear Additive Smooth Structures and Complicated Models for Large Scale Medical Data

Brian D. Marx (1), Paul H. C. Eilers (2), and Dorothee Auer (3)
Department of Experimental Statistics, Louisiana State University (1)
Department of Medical Statistics, Leiden University, Netherlands (2)
Max Planck Institute, Germany (3)

OBJECTIVES: Modern medical instruments can generate extensive data, which present unique problems when building models. Our motivating example is a study on tumors where the data consist of MRI signals, histograms, covariates with possibly non-linear influences, and standard linear regressors and factors. To meet this challenge we apply a modeling approach in which generalized additive models, regression on signals, varying coefficient regression and standard regression are combined in a unified way. Estimation is achieved through penalized likelihood, avoiding backfitting and knot optimization schemes. Fast cross-validation and the computation of standard errors and regression diagnostics present no problems.

METHODS: The model can contain linear covariates (L), any combination of smooth additive components (G), and high dimensional signal components (S); additionally varying coefficient components (V) are allowed. We assume that G is, and the coefficients of V and S are, inherently smooth -- projecting each of these onto B-spline bases using a modest numbers of equally spaced knots. Further smoothness is achieved through difference penalties on adjacent B-spline coefficients (the P-spline approach). The linear re-expression allows simultaneous estimation of all components. We regulate the flexibility of each component through an independent penalty parameter, optimally chosen based on cross-validation or information criteria. Additionally, since we stay close to the (generalized) linear model framework, diagnostics are tractable. As the model boils down to a large penalized generalized linear model, the dependent variable can be metrical, binomial or Poisson.

RESULTS: We show how to combine B-splines, discrete penalties, and GLM ideas to construct a relatively all-purpose, fast and compact regression technique. We illustrate the methods using metrical, Poisson, and binary response variables. Rather complex data structures are modeled as regressors, for example functional regressors in the form of MRI signals, constructed histograms of diffusion data for each patient, smooth functions of age, varying coefficients for seasonal impacts and other, more traditional, covariates. The computational details are transparent to the user, although interesting. The gam() function in S-PLUS is coerced to handle these structures simultaneously.

CONCLUSION: We think that we have an extremely practical solution to a complex problem. Although some of the ideas presented are addressed in the existing literature, a practical solution is lacking, especially when addressing all of these structures simultaneously. The medical examples with complicated data illustrate the usefulness of the methodology.

Contributed Session 3C
Tuesday, January 23, 2001
2:00 – 3:30

Multiple Comparison with a Control in the Presence of Correlated Data

Jean G. Orelien (1), Richard Morris (1), and Chris Gotwalt (1,2)
Analytical Sciences, Inc. (1)
North Carolina State University (2)

OBJECTIVES: Researchers in toxicology are often interested in comparing several doses of a xenobiotic agent versus a control. Methods for multiple comparisons (MCPs) that could be used to perform these statistical tests were developed for traditional analysis of variance (ANOVA) models where the observations can be assumed to be independent. The assumption of independence, however, may not be appropriate for data from developmental toxicity studies in which measurements are taken on littermates. Although a variety of approaches to analyzing correlated data, including methods of generalized estimating equations (GEE), have been proposed, methods for simultaneous inference necessary to identify which dose levels have a significant effect with a given experiment wise alpha level have not been discussed for correlated data in toxicology experiments.

Objective: In this paper, we present methods for performing MCPs with a control in generalized linear models. These methods were developed to analyze data from from the National Toxicology Program (NTP).

METHODS: We show that the p-values for the test can be obtained by computing a probability from a multivariate normal whose variance is a correlation matrix. Computation of a multivariate normal probability can be computer intensive and may not be readily available from current statistical software packages. However, an algorithm presented in Genz and Bretz (1999) can be used to compute directly the multivariate normal probability involved in the computation of the p-values. By assuming that the variance of the multivariate normal has a product factorial structure, an algorithm by Hsu (1992) can also be used to approximate these p-values. Joint confidence intervals for the difference between treatments and control can be obtained by inverting the test. Previously published data will be used as an example to illustrate the proposed method. To assess the validity of the test, we simulated correlated data on 2 endpoints (one continuous, the other dichotomous) from the NTP so that the type I error rate and power could be observed.

RESULTS: The simulation show that this test provides adequate experiment wise type I error rates and reasonable power. The power is greater or close to 80% when the correlation among pups from the same litter in the control group is no more than 0.2.

CONCLUSION: We recommend using this test to perform multiple comparisons with a control in generalized linear models. This test can also be extended to perform multiple comparisons when a different method other than GEE such as mixed model is used to account for the correlation structure among littermates. The only requirement is that the parameter estimates have a normal or asymptotic normal distribution.

Impact of Ignoring Centers on Evaluating the Influence of Process Measures

Laura P. Coombs (1), Elizabeth R. DeLong (2), and Eric D. Peterson (2)
Duke Clinical Research Institute (1)
Duke University Medical Center (2)

OBJECTIVES: Surgical quality improvement initiatives attempt to derive standards of care based on the demonstrated effects of various process measures. In evaluating the influence of such medical processes on surgical outcomes, the "center effect" (i.e., the impact of differences among centers on outcomes of interest) represents an important, sometimes over-looked consideration. Unlike patient-specific risk factors, which have similar effects on surgical mortality across a broad population, the tendency to use certain processes varies from one center to another. The popular use of propensity scores to control for differences in patient constituency between comparison groups will not account for differences in outcome due to differences among centers. The objective of this paper is to demonstrate the importance of considering the center effect in the analysis when evaluating the effectiveness of a process, regardless of whether risk adjustment or propensity scores are used.

METHODS: Selecting data contributed by nine centers from 1996-1999, we examined the effects of two processes on operative mortality for patients undergoing isolated Coronary Artery Bypass Grafting (CABG). Process A is a process with standardized protocols and patterns of use across centers, whereas usage for Process B ranged from 23.9% to 85.7%. Initially, we used two methods to examine the effects of each process, both ignoring the center effect. First, a logistic regression model was used to determine the effect of the process after adjusting for patient risk factors. Second, propensity scores for each patient were developed and the effect of the process was examined among subgroups of patients with similar propensity. Next, we modified these two methods to include center. Center was added to the risk adjustment model (as a random effect) and to the propensity model. The effect of the process within subgroups based on overall propensity score was examined as well as the effect for subgroups based on propensity score within center.

RESULTS: The effect of Process A, which had relatively standard patterns of use across centers, remained stable regardless of the analysis used. The effect of Process B, however, was reduced considerably after adjusting for center effect. Ignoring center effect, the OR for Process B after adjusting for risk factors was 0.66 ($p < .01$). Using a mixed model to adjust for center effect, the OR for Process B increased to .83 ($p = .04$). For the propensity score analysis, the average reduction in mortality across subgroups was 1.2% when center was not included in the propensity model. Adding center, the average reduction in mortality dropped to 0.4% and remained at 0.4% when using propensity within center to stratify.

CONCLUSION: It is important to include centers in the analysis of a process effect, especially when use of that process varies across centers. Ignoring the center effect in the development of a propensity score or risk-adjustment model may produce misleading results.

Using 'Lowess' to Remove Systematic Trends from Contaminated Data Prior to Logistic Regression with Empirical Quantile-Categories: A Case Study Involving Esophageal Cancer.

Craig B. Borkowf, Paul S. Albert, and Christian C. Abnet.
Cancer Prevention Studies Branch, Division of Clinical Sciences, National Cancer Institute

OBJECTIVES: To demonstrate the utility of locally weighted robust regression ('lowess') as a method for estimating and then removing systematic trends-over-time from data; and to apply this method to a nested case-control study of sphingolipids as predictors of esophageal cancer.

METHODS: We developed this methodology in response to the statistical challenges in the following study. Epidemiological theory suggests that the consumption of corn and wheat infected by fungal toxins, such as those in the fumonisin family which disrupt the normal sphingolipid pathways, increases the risk of esophageal cancer. To test this hypothesis, about 100 cases and 200 controls (stratified on age and gender) were selected from the participants in an extensive nutrition intervention trial conducted in China. These subjects' blood serum levels of the sphingolipids sphinganine (Sa) and sphingosine (So) were measured as biomarkers for fungal toxin exposure. These sphingolipid levels were measured by HPLC analysis over a time period of about 60 days, during which adjustments were made to the HPLC machine. Unfortunately, when the Sa measurements were plotted against day of HPLC analysis, the mean levels for the cases and controls showed systematic trends-over-time. The lowess estimates of these two trends had approximately the same shape. Because these serum samples are a valuable and limited resource, we decided to estimate and then remove the common trend by lowess. We then performed logistic regression with case/control status as the response and several predictors, including the lowess adjusted Sa measurements categorized by the empirical quartiles of the controls. Ideally, lowess will recover the relative differences between the continuous Sa measurements, which is sufficient for logistic regression with empirical quantile-categories. To validate this trend removal method, we conducted extensive simulations with various models for the distribution of the cases and controls and with various shapes and relative magnitudes of the contaminating trend-over-time.

RESULTS: Logistic regression with empirical quantile-categories computed from the lowess adjusted Sa data suggests that the risk of esophageal cancer does not change over the quartile-categories of Sa. This result is consistent with that for the So measurements, which were not contaminated. More generally, our simulations show that lowess can be successfully used to estimate and remove systematic trends-over-time prior to logistic regression with empirical quartile-categories. Using the trend contaminated data tends to give attenuated parameter estimates and hypothesis tests with subnominal significance levels and low power. Conversely, using the lowess adjusted data tends to give nearly unbiased parameter estimates and hypothesis tests with nearly nominal significance levels and improved power.

CONCLUSION: We conclude that under minimal assumptions lowess can be used to remove systematic trends-over-time from data, and that logistic regression with empirical quantile-categories computed from the lowess adjusted data gives approximately correct results. We used this method of analysis to conclude that the risk of esophageal cancer does not change over the quartile-categories of Sa.

Measuring Disease Progression with Clinical Risk Groups

Jon Eisenhandler (1), Norbert I. Goldfield (1), Richard F. Averill (1), and John H. Muldoon (2)
3M Health Information Systems (1)
National Association of Children's Hospitals and Related Institutions (2)

OBJECTIVES: One of the problems confronting policy makers and physicians is the inability to accurately track disease progression in large, disparate populations. There is no reason this should be the case. Considerable data, albeit of varying quality and completeness, are captured in the routine processing of claims. Moreover, the increased automation of claims submission, the move to computerized medical records, and the increasing inclusion of previously unavailable data such as lab results are likely to increase the amount of accessible and useful data in the foreseeable future.

Over the last decade, a number of risk adjustment systems have been created. Using routine claims data, these systems work by assigning an individual to a group or groups reflecting the individual's clinical and other characteristics. This assignment is then used to predict future resource utilization or analyze past resource utilization. Clinical Risk Groups (CRGs) is such a system. It is different from other risk adjustment systems in that it is a categorical model with explicit severity of illness levels. Each individual is assigned to a single group which encompasses their whole medical history. This group reflects the individual's chronic diagnosis or diagnoses and their severity.

If CRGs are assigned at specified intervals, changes in CRG assignment and therefore individual health status can be identified. When aggregated, this will produce observable trends in disease progression which can be linked back to clinical practices and other factors.

METHODS: Using Medicare data for 1992 - 1994, a subpopulation of individuals with standard Part A and Part B coverage for all three years have been identified along with all of their Medicare claims. CRGs are assigned for each year. The assignments use data from single years and data from multiple years. A set of CRG defined cohorts for common diseases, diabetes mellitus, emphysema, and hypertension at low levels of severity are identified for the base year. Movement from the cohorts over time are tracked and described.

RESULTS: There is significant movement between CRGs over time. Most of the movement is predictable with patients with chronic conditions having stable or worsening health. Some patients, however, become healthier. Disease progression is affected by age and sex.

CONCLUSION: CRGs are able to identify changes in health status over time. These changes can be linked to demographic factors. There is no reason they can not also be linked to clinical practices and to the physicians or physician groups responsible for clinical practices.